View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 96 (2013) 512 – 521

Procedia
Social and Behavioral Sciences

13th COTA International Conference of Transportation Professionals (CICTP 2013)

Modeling commuters' travel behavior by Bayesian networks

Jingxian Wu[a*], Min Yang[a]

[a] *Transportation college of Southeast University, Sipailou 2#, Nanjing,* 210096*, China*

Abstract

Previous studies indicate that residential location and commute distance may influence individual's travel behavior, but most models are limited to capture the internal relationship. In this paper, the methodology of Bayesian networks is introduced. With the combination of network structure and conditional probability table, Bayesian networks are capable of capturing the uncertainty nature. Moreover, the data is based on Fuyang Resident Travel Survey in 2012. The outputs illustrate that commute distance influences commuters' travel mode choice directly while the residential location doesn't. The experimental results show that different groups of people have different reactions to distance on mode.

© 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of Chinese Overseas Transportation Association (COTA).

Keywords:Bayesian networks, travel behavior, commute distance,residential location

1. Introduction

The majority of research on travel behavior has focused on limited information such as individual and trip attributes, while a number of other factors such as residential location and commute distance have been neglected. With the advancements of this field, many researchers started to reconsider the impacts of residential location characteristics on travel behavior characteristics such as trip distance (Naess, 2006), mode choice (Van, Holwerda and Van, 2004; Zhang, 2004; Pinjari, Pendyala, Bhat and Waddell, 2011), auto ownership (Pinjari, Pendyala, Bhat and Waddell, 2011; Bhat, Guo, 2007) and so on. Kenneth et al (2008) investigated the influences of the built environment and transport availability on travel behavior on the dataset of in the South Bay region of Los Angeles County, California (Joh, Boarnet, Nguyen, Fulton, Siembab and Weaver, 2008). Xinyu et al (2007) have found that residents living in the traditional environment tend to drive less than those living in the suburban neighborhoods (Cao, Mokhtarian and Handy, 2007; Bagley, Mokhtarian, 2002). Frank et al (2006) commented the framework of travel cost in the respects of land use and urban models, and explained that the urban modes would affect commuters' stopping times directly. Additionally, they have found some difficulties to figure out the impacts of residential location (Frank, Bradley, Kavage and Chapman, 2008). Some analyses assumed the residential

* Jingxian Wu. Tel.: 15062254598;
E-mail address: 624107036@qq.com

1877-0428 © 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of Chinese Overseas Transportation Association (COTA).
doi:10.1016/j.sbspro.2013.08.060

variables to be exogenous variables through utility maximizing models which ignored the unobserved relations between the variables and residential location (Hang, 2012). Most of these models are difficult to handle the relationship between the dependent and independent variables, which is an obstacle for new variables to add in. Some researchers adopted the neural network to solve this problem, but they found that the network structure is not easy for understanding as it can't explicitly represent mechanism of travel behaviors. The comparable model is about the structural equations, with the variable increasing the structure of which is hard to understand. Therefore, the primary aim of this paper is to apply a new approach for solving the difficulties above and in-depth understanding of the different regional commuters' travel mechanism, by which the reasonable regulations or policies can be made to lead the construction of public transport. So the new method should include these properties as followed:

- It will be able to incorporate new elements into the model without confinements
- Then it can capture and represent the potential relationships among the variables jointly.
- It must own the function to explain the travel mechanism conveniently for better understanding and further analysis.

Attempting to answer these questions, numerous methods were compared to find that Bayesian networks are the optimal choice. A BN, which is based on a process of inductive knowledge discovery, is a method that combines Bayesian probability theory and graph theory, and it has been widely used in the overlapping fields such as artificial intelligence and machine learning (Heckerman, Geiger, Chickering, 1995). BNs are based on the dependence among variables to identify potential causal dependencies among variables. Consequently, Bayesian networks make it easier for researchers to add new variables. Moreover, Pearl (1994) indicated that the causal information encoded in BNs and the output of the BNs could facilitate the analysis of situations. Janssens et al (2003, 2006) suggested that Bayesian networks are better suited to capture the complexity of the underlying decision-making. Their findings mean that: 1) BNs are able to integrate the residential location into the model, 2) The uncertainty relationships among the variables can be captured and also be visually presented in the form of network structure and conditional probability tables, 3) the travel mechanism revealed in the model seems more reasonable.

This paper utilizes the 2313 trip samples of 2012 residential travel survey to assist the analysis of the relative associations between residential location and commuters' travel behavior. The 2012 Resident Travel Survey was conducted in a Chinese city Fuyang in the form of questionnaire, in which attributes related to household-related variables (address, type, age, and area about the house), individual variables (gender, career, age, the ownership of a license, education and salary etc.) and the trips were recorded. Furthermore, the local authorities also supplied some materials on geography characteristics of the city by which the residential location attribute was concluded.

The structure of the paper is organized as follows. Section 2 gives an overview of the data that is used and a brief flow-chat of the modeling procedure. Section 3 elaborates on commuters' travel behavior model as well as the application of a Bayesian network, and explains them in details. Section 4 provides an accuracy evaluation of the outputs. Finally, conclusions are reported.

2. Approach and data

This research utilizes the model of BNs to deliberate the mechanism of residential location and commute distance towards travel mode choice through the analysis of impacts of individual attributes, household-related attributes, travel characteristics and the residential location. A flowchart is supplied to illustrate the procedures of the paper in figure1. The processing flow has three steps: 1) establishing a database for the model's preparation, 2) modeling the commuters' travel behavior and 3) analyzing the output of the model.

Before the procedure of modeling, the background of the city Fuyang is introduced first. It is a densely populated city which is located in Anhui province of China. Based on the administrative boundary and

geographical features the city is classified into three districts: Yingquan, Yingdong and Yingzhou. They are naturally split by the Ying River from northwest to southeast and Quan River from west to east (refer to figure 2). Yingzhou is featured as the political center which assembles education, entertainment industry, health center and so on. Yingquan is a commercial center along with high quality residential area and Yingdong is featured as the labor-intensive industry center. It has long been known that different locations result in different kinds of characteristics for each zone and people in the same districts tend to form somewhat similar habit.

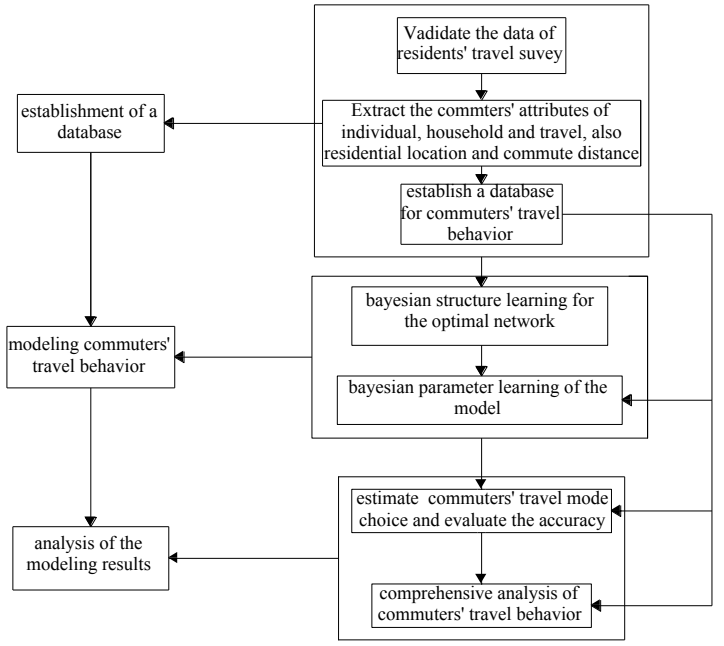

Figure 1 Flow chart of the modeling procedure

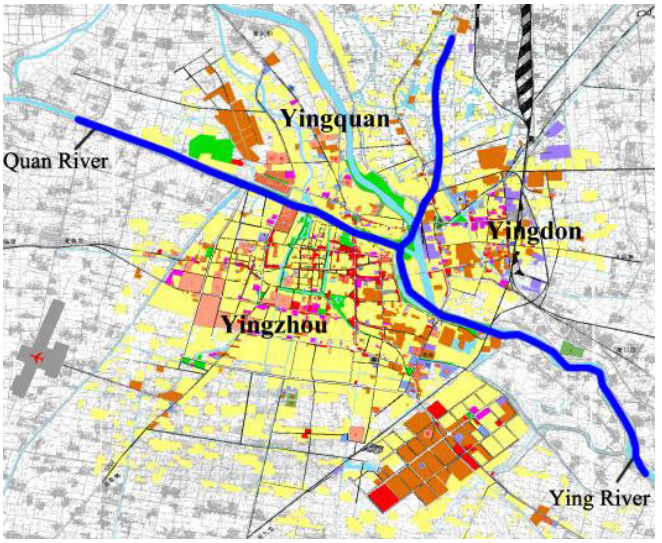

Figure 2 Geographic locations of the three districts

In the procedure of database establishment, the main thing focuses on how to organize the related information from the validated data so that they can be used as the input for the modeling. Based on the research, the data is screened and merged into three categories. In many cases, the number of classes within each variable has been reduced to simplify the analysis. For instance, the variable "travel mode" for the respondents has been reduced to just three classes: public bus, electric bicycle and private car. The variable "career" has been reclassified into three categories and the "age" is aggregated to four classes ranging from 20 to 59 years. Then it comes to the commute distance which is converted from the commute time, and from the basic materials of the city transport in 2007 the regular speeds of the electric bicycle and the private car are 20 km/h and 35 km/h. In order to get the average speed of the public buses a comprehensive inspect was organized in 2012, and the result showed the speed was about 17km/h. A notable point is that the commute time which is obtained from the starting and reaching time is not accurate, and it needs to be corrected by subtracting the 5 minutes' waiting time. Multiplying the speed responding to the mode and the revised commute time hence gets the distance. The variable "peak" is determined by the travel time distribution of the city Fuyang from the statistics and the peak hours are from 7:00 to 9:00 and from 17:00 to 18:00. The residential location is derived from the address line from the data. In the preliminary screening 16 variables will be chosen for the construction of a Bayesian network in the section 3. The information for these variables used in the study and their detailed classification is referring to Table1:

Table1 Nodes variables and their values

attribute	variable	description
Household attributes	workers	Number of workers, 1 to 6
	e-bicycles	Number of electric bicycles, 0 to 5
	cars	Number of cars, 0 to 3
	prechild	Number of preschool children, 0 to 2
	hsize	Number of Household members, 1 to 6 or more
Individual attributes	location	Residential location: 1-Yingzhou, 2-Yingquan and 3-Yingdong
	gender	gender of a commuter: 0-female, 1-male
	career	Career type: 1-administrative staff, 2-company employee, 3-private entrepreneur
	license	Driver status of respondent: 0-not licensed, 1-licensed
	age	Age of respondent, 1-20 to 29years, 2-30 to 39years, 3-40 to 49years, 4-50 to 59years
	role	The role in the household, 1-child, 2-wife, 3-husband, 4-grandparents or 5-other
	salary	Personal salary in 4 classes, 1-less than 1000yuan, 2-1000 to 2500yuan, 3-2500 to 4500yuan or 4-more
	education	Education level, 1-middle school and below, 2-high school, 3-undergraduate and above
Trip attributes	peak	Starting time of trip at peak hour or not: 0-no, 1-yes
	mode	Travel mode, e-bicycle, bus, car
	distance	Kilometers of commute distance, 0-20kilometers divided by "mode" distribution characteristics on "distance"

3. Modeling of travel behavior

A Bayesian network, which is also known as belief networks and Bayesian belief networks, A Bayesian network can be described as a directed acyclic graph (DAG) and a set of conditional probability tables (CPT). In the DAG G=(X, E), $X=\{x_1,\cdots, x_n\}$ denotes the set of nodes corresponding to random variables . The arc between two nodes represents a causal relation: the node from original is called the parent node and the other node is called the child node (*14*), so the variables in Π_i are the parents of the node corresponding to x_i. Associated with each node x_i is one conditional probability table which holds conditional probability

distributions $p(x_i \mid \Pi_1)$. According to the Bayesian probability theory the joint probability distribution of all nodes is factorized as in

$$p(x_1,...,x_n) = \prod_{i=1}^{n} p(x_i \mid \Pi_i) \quad (1)$$

The output of a BN does not only reveal the qualitative relationships between the attributes of travel, but also supplies the quantitative measures, in the form of conditional probability distributions, of how the various factors affect and interact with each other. Using a BN for analysis of the residential travel data can uncover and characterize the interaction the residential location and travel behavior. Therefore, the process of obtaining a BN from the data mainly involves two steps: structure learning and parameters learning.

3.1. Structure learning

The aim of structure learning is to search for the optimal topology that presents the associations among the variables, in other words. A structure of network can be obtained by the priori domain knowledge and training dataset (Jiawei, Kamber, 2006). Prior studies indicate that the former option is applied to identify the nodes in the structure while the latter one which is a pure data-driven method can be adaptively employed for structure and parameters learning (Xiujun, 2002). The approach used for structure learning can be divided in two categories: search-and-score and constraint-based. The constraint-based approach is used less often as the repeated independence tests lose statistical power (Marco, Scuderi and Kelly, 2005). Many algorithms used in the search-and-score approach sometimes may link unnecessary variables in complex way which will decrease the efficiency. Markov Blanket learning algorithm, based on a heuristic search, can make up the deficiency. Markov Blanket learning, denoted as MB, is an efficient algorithm for the BN structure entirely dedicated to the characterization of the target variable "mode" (Shuangcheng, Sengseng and Hui, 2004; Aliferis, Tsamardinos and Statnikov, 2003). In the procedure of this paper, some unconcerned variable nodes and links will be deleted by MB to simplify the structure. At the same time the efficiency the modeling is improved obviously without too much decrease in accuracy. Combining the two methods gets a network as figure 3.

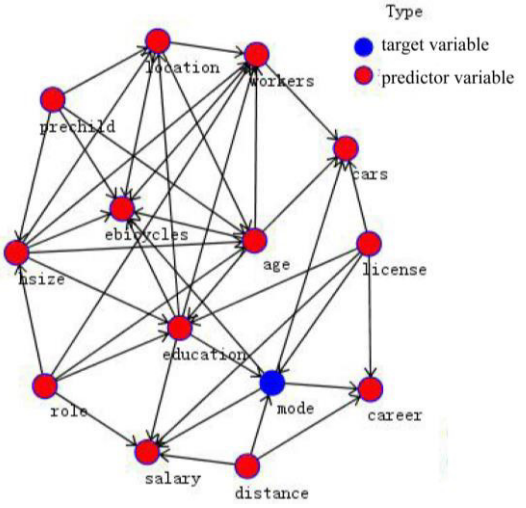

Figure 3 Bayesian network of commuters' travel behavior

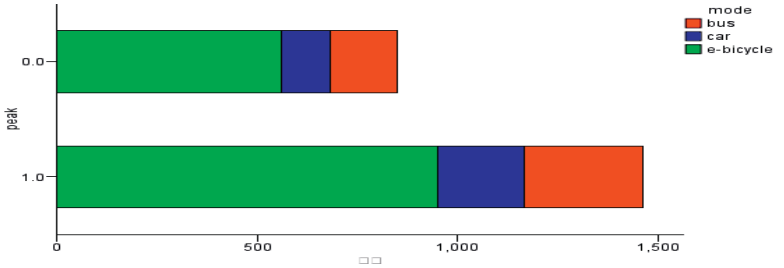

Figure 4 The mode choice and peak

In the network, the number of variables is reduced from 16 to 14. It can be seen that variables of peak and gender are screened out of the dataset. To test the results, the paper analyzed the two variables and mode choice respectively. For instance, the figure 4 shows that the two periods of variable "peak" have a similar mode choice distribution which attributes to the inelasticity of the commuters' travel, so there is no evident association between them. The residential location has an indirect association with the mode choice, and the mechanism of this relation is displayed in the network. It indicates that the residential location has some influences on the commuters' travel mode choice to some extent. The variable location is influenced by the variables of education, hsize and prechild. In China, the housing price of the urban is more expensive than that of the suburban, and the urban is well-developed and crowed. Therefore, people with a high education have more opportunities to live in the urban like Yingzhou district. People with children are more inclined to live in the urban, because the urban area has better education, health care systems and so on. Since the price of house is high and space of the urban is crowed, the expanded families sometimes have to live in the suburban. These are interpretations about the figure based on general knowledge of real life. These are all perceptions about the picture, but it needs some data to look into the associations of them. Therefore, the following part is about parameters learning for conditional probability tables.

3.2. Parameter learning

With the constructed network and dataset parameter learning can proceed for specifying the probability distribution for each node. As the data supplied by the survey is complete, there are two terms about the parameters learning: Maximum Likelihood Estimation and Bayesian. The former approach, based on the concept of traditional statistical analysis, evaluates the goodness of fit on samples and parameters according to the likelihood of samples and model. MLE has the advantages of consistency, incremental effectiveness and flexibility and its likelihood function goes as follow:

$$L(\theta; D) = p(D|\theta) = \prod_i p(x_i|\theta) \tag{2}$$

CPT of each node is achieved by utilizing the SPSS Clementine software and the survey data. Two examples are given in table 2 and table 3.

The table 2 shows that, workers with no driving license and a lower education are tend to drive an e-bicycle for work even the distance increasing while those with licenses and high education levels shift mode choices from bus and e-bicycle to private car. It indicates that the latter group of people is more sensitive to the commute distance. When the distance exceeds the value 16.116 kilometers most people will be more inclined to the mode car except those without license and who have received lower education. From these data further analysis can be carried out by statistic tools. In order to facilitate comparative analysis, two figures are derived by the information in table 2. As the figure 5 shows, people with no license but higher education use the bus more frequently than those with lower one. Compared with the other two lines, the tendency of blue line is relatively stable with the increase distance. To a certain extent, it can be explained that when the commute distance is within 5km or ranging from 12km to 16km, highly educated people prefer taking bus to work. In the next picture, there is no significant difference between the three groups in the respect of the car use. Licensed one would rather take their car with the increasing distance. Those highly

educated commuters use the car slightly more than others. When the commute distance reaches 16km, all of them will turn to the car for work. However, when it comes to 12km, some of them refuse the car use that can be attributed to the high travel cost, and they shift the other two modes.

Table 2 Conditional probabilities of mode

parents			probability		
license	**education**	**distance**	**bus**	**car**	**e-bicycle**
0	1	<=4.466	0.128	0	0.871
0	1	4.466~8.35	0.083	0.016	0.899
0	1	8.35~12.233	0.066	0.009	0.924
0	1	12.233~16.116	0.142	0	0.857
0	1	>16.116	0	0	1
0	2	<=4.466	0.394	0.005	0.6
0	2	4.466~8.35	0.177	0.002	0.819
0	2	8.35~12.233	0.134	0	0.865
0	2	12.233~16.116	0.432	0.027	0.54
0	2	>16.116	0	0.111	0.888
0	3	<=4.466	0.513	0	0.486
0	3	4.466~8.35	0.289	0.014	0.696
0	3	8.35~12.233	0.246	0.038	0.714
0	3	12.233~16.116	0.59	0	0.409
0	3	>16.116	0.333	0.666	0
1	1	<=4.466	0.384	0.076	0.538
1	1	4.466~8.35	0.142	0.171	0.685
1	1	8.35~12.233	0.043	0.608	0.347
1	1	12.233~16.116	0	0	1
1	1	>16.116	0	0.75	0.25
1	2	<=4.466	0.34	0.021	0.638
1	2	4.466~8.35	0.189	0.153	0.657
1	2	8.35~12.233	0.024	0.59	0.385
1	2	12.233~16.116	0.3	0.2	0.5
1	2	>16.116	0	0.972	0.027
1	3	<=4.466	0.403	0.112	0.483
1	3	4.466~8.35	0.191	0.353	0.454
1	3	8.35~12.233	0.046	0.695	0.257
1	3	12.233~16.116	0.2	0.68	0.12
1	3	>16.116	0	0.953	0.046

Table 3 Conditional probabilities of location

parents			probability		
education	**hsize**	**prechild**	**Yingzhou**	**Yingquan**	**Yingdong**
1	1	0	1	0	0
1	2	0	0.558	0.238	0.205
1	2	1	1	0	0
1	3	0	0.572	0.139	0.288
1	3	1	0.594	0.164	0.24
1	3	2	1	0	0
1	4	0	0.566	0.366	0.066
1	4	1	0.5	0.441	0.058
1	4	2	0.5	0.5	0

1	5	0	0.421	0.526	0.052
1	5	1	0.259	0.555	0.185
1	6	0	1	0	0
1	6	1	0.222	0.777	0
1	7	0	0.5	0	0.5
1	7	1	0.8	0.2	0
1	9	2	0	1	0
2	1	0	0.5	0	0.5
2	1	1	0	0.333	0.666
2	2	0	0.581	0.22	0.197
2	2	1	0.8	0	0.2
2	3	0	0.662	0.132	0.205
2	3	1	0.701	0.155	0.142
2	3	2	1	0	0
2	4	0	0.52	0.265	0.214
2	4	1	0.515	0.272	0.212
2	4	2	0.166	0.833	0
2	5	0	0.833	0.111	0.055
2	5	1	0.552	0.368	0.078
2	5	2	0	0	1
2	6	0	0.25	0	0.75
2	6	1	0.818	0	0.181
2	6	2	1	0	0
2	7	0	1	0	0
2	7	1	0.4	0.6	0
3	1	0	0.714	0	0.285
3	1	1	1	0	0
3	2	0	0.745	0.096	0.157
3	2	1	1	0	0
3	3	0	0.871	0.087	0.041
3	3	1	0.794	0.123	0.082
3	3	2	0.5	0	0.5
3	4	0	0.883	0.069	0.016
3	4	1	0.764	0	0.235
3	5	0	0.76	0.16	0.08
3	5	1	0.76	0.16	0.08
3	5	2	0	0	1
3	6	1	1	0	0
3	6	2	1	0	0

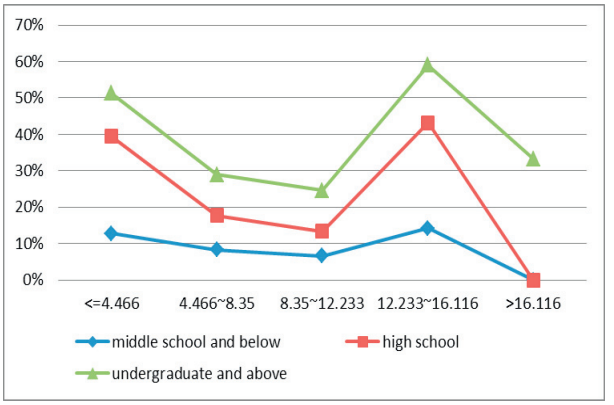

Figure 5 Bus usage of people without license

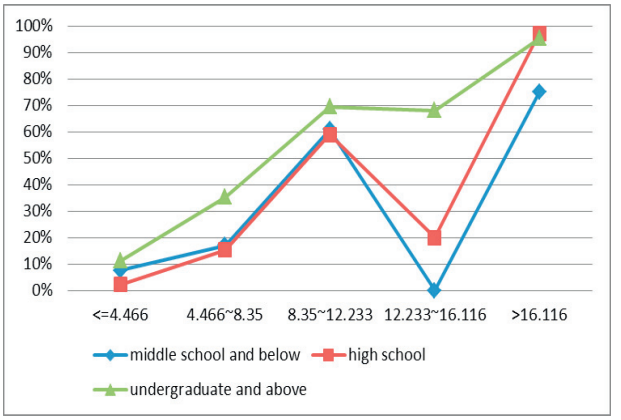

Figure 6 Car usage of people with license

3.3. Evaluation

Obviously, accuracy is the most important evaluation criterion of any model. In order to know the accuracy of model, the first and most important thing is to predict the values of target "mode". The list of individual trip data conducts as the input to the model then their travel mode choice results can be achieved. If the modeling results are consistent with the observation choice, it means "hit". . The model is validated by survey data. The hit rate calculated by the above method is 90.23% meaning the model of better practicability.

4. Conclusion

This paper constructed a model of the commuters' travel behavior by the Bayesian Network, in which the elements of residential location and commute distance were incorporated, and it was developed on the basis of 2012 Resident Travel Survey in Fuyang, China. The structure, which presents the travel behavior, was obtained by the Markov Blanket learning algorithm, while the CPT of each node was achieved through the method of Maximum Likelihood Estimation. The accuracy estimation showed the BN model was effective for prediction and analysis.

The outputs of the modeling illustrates that the commute distance has a direct influence on commuters' travel mode choice while the residential location has an indirect effect on it. Though the CPT of "mode" the

authors find that the influence degree of distance on mode choice is widely diverse from different people. In the future, we hope the distance data can be obtained through an accurate way and more detailed materials of the residential location can be supplied for different studies.

Acknowledgements

This research is supported by National Basic Research 973 Program (2012CB725400) and National Natural Science Foundation of China (50908052, 51178109, 51008061, 51108080). Fundamental Research Funds for the Central Universities and Foundation for Young Key Teachers of Southeast University are also appreciated.

References

Naess P. Accessibility, activity participation and location of activities: exploring the links between residential location and travel behavior. Urban studies, Vol.43, No.3, 2006, pp.627-360
Van Wee B., Holwerda H., Van Baren R. Preferences for modes, residential location and travel behaviour: the relevance for land-use impacts on mobility. European Journal of Transport and Infrastructure Research, Vol.2, No.3/4, 2004, pp.305-316
Zhang M. The role of land use in travel mode choice: evidence from Boston and Hong Kong. Journal of the American Planning Association, Vol.70, No.3, 2004, pp.344-360
Pinjari A. R., Pendyala R. M., Bhat C. R., and Waddell P. A. Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. Transportation, Vol.38, No.6, 2011, pp.933-958
Bhat C. R., Guo J. Y. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. Transportation Research Part B: Methodological. Vol.41, No.5, 2007, pp.506-526
Joh K., Boarnet M. G., Nguyen, M.T., Fulton W., Siembab W., and Weaver, S. Accessibility, travel behavior, and New Urbanism: Case study of mixed-use centers and auto-oriented corridors in the South Bay region of Los Angeles, California. Transportation Research Record: Journal of the Transportation Research Board, Vol.2082, No.1, 2008, pp.81-89
Cao X, Mokhtarian P. L., and Handy S. L. Do changes in neighborhood characteristics lead to changes in travel behavior? A structural equations modeling approach. Transportation, Vol.34, No.2, 2007, pp.535-556
Bagley M. N., Mokhtarian P. L. The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. The Annals of Regional Science, Vol.36, No.2, 2002, pp.279-297
Frank L., Bradley M., Kavage S. and Chapman. J. Urban form, travel time, and cost relationships with tour complexity and mode choice. Transportation, Vol.35, No.1, 2008, pp.37–54
Pearl J. A probabilistic calculus of actions. Proceedings of the Tenth international conference on Uncertainty in artificial intelligence, 1994, pp:454-462
Janssens D., Wets G., Brijs T., Vanhoof K., Timmermans H. Identifying behavioral principles underlying activity patterns by means of Bayesian networks. Electronic conference proceedings of the 82nd Annual Meeting of the Transportation Research Board, 2003, pp:12-16
Janssens D., Wets G., Brijs T., Vanhoof K., Timmermans H., Arentze, T.A. Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. European Journal of operational research, Vol.175, No.1, 2006, pp.16–34
Heckerman D., Geiger D., Chickering D. M. Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning, Vol.29, No.3, 1995, pp.197–343
Neapolitan R. E. Learning Bayesian networks. Pearson Prentice Hall Upper Saddle River, New Jersey, 2004
Jiawei H., Kamber M. Data Mining concepts and techniques, China machine press, China, 2006
Xiujun G. Research on Bayesian learning theory and its application, Doctoral Dissertations of institute of computing technology Chinese academy of science, Beijing, 2002
Shuangcheng W., Sengseng Y. and Hui W. learning Markov Blanket prediction based on Bayesian networks. Pattern recognition and Artificial intelligence, Vol.17, No.1, 2004, pp.17–21
Hang L. the influence of residential location and commute distance on travel mode choice, Bachelor's Thesis of southeast university, Nanjing, 2012
Marco G. Scuderi and Kelly J. Clifton. Bayesian approaches to learning from data: using NHTS data for the analysis of land use and travel behavior. Journal of transportation and statistics, Vol.8, No.3, 2005, pp.17–21
Aliferis C. F., Tsamardinos I., Statnikov A. HITON: A novel Markov Blanket algorithm for optimal variable selection. AMIA Annual Symposium Proceedings, 2003, pp: 21-25